基于微博的电影首映周票房预测建模*

王晓耘　袁　媛　史玲玲
(杭州电子科技大学管理学院　杭州 310012)

摘要:【目的】解决现有的票房预测模型由于数据受限等因素导致的无法实现在影片上映前进行票房预测这一问题。【方法】在获取微博评论的基础上, 使用 SVM 识别出消费者的显式消费意图, 即强正面评论; 对传统的分类准则进行修正, 构建基于 HowNet 的中文微博情感词典, 进而定义一个新的用户影响力特征; 使用 BP 神经网络进行票房预测。【结果】实验结果表明, 本文建立的模型能够较为准确地对电影首映周票房进行预测。【局限】由于语料不充分, 本文构建的中文微博情感词典, 可能会无法在所有的电影微博评论中表现出较好的分类效果; 此外也没有建立一个能够在电影上映周期内动态预测票房的票房预测模型。【结论】该模型能够有效地进行首映周票房预测, 具有现实的可行意义。

关键词: 情感词典　情感分类　首映周票房预测　神经网络
分类号: TP391.1　G35

1　引　言

随着微博的快速发展, 不断增加的在线评论正在极大地影响着传统的市场决策, 使得文本挖掘成为商业界与学术界共同的热点话题[1]。而其中将微博与传统事件预测相结合的研究成为一大热点[2]。如 Liu 等[3]在使用贝叶斯分类器对博客进行情感分类的基础上进行票房预测, 预测结果可以为电影上映期间的银幕分配或营销支出提供数据支持。由于电影制作和营销推广需要大量投资又有极高的风险, 因此电影产业成为预测和市场评估的重点领域[4-9]。但我国电影行业始终面临着由于票房预测方法和工具缺失导致的投资者无法有效对冲投资风险这一问题。事实上, 作为仅次于美国和日本的全球第三大电影生产国, 我国目前只有极少数电影投资是盈利的, 70%的国产电影基本都难以回收成本[5]。微博作为娱乐行业的影响力阵地, 电影主演发布的每一条和电影相关的微博都能够引发一系列评论和转发进而转化为相应的消费行为, 如董成鹏在 2015 年 7 月 16 日发布的一条和《煎饼侠》相关的微博, 引发了 3 179 条转发和 5 191 条评论。同时用户的每一条评论所汇聚成的集体智慧, 都能体现出电影主创人员的票房号召力, 都会蝴蝶效应般地影响实际票房。因此, 在这一背景下利用微博评论进行票房预测无疑对电影生产商来说具有重大的现实意义。

2　相关研究

2.1　基于微博或博客的票房预测

为了充分挖掘微博中的信息进行票房预测, 国内外学者进行各种探索。常用的方法是利用微博中的转发数、粉丝数、评论情感作为模型的输入, 采用情感自回归模型、神经网络或 SVM 等进行票房预测。

张闯等[2]将节点属性分为动态和静态两类, 提出基于节点属性进行信息预测的 AVN 模型, 并使用 BP 神经网络进行票房预测。但该模型对情感倾向的定义忽略了不同的情感对票房的贡献能力。Liu 等[3]首次将博客中的情感看成是一系列因素共同作用的结果, 提

通讯作者: 袁媛, ORCID: 0000-0003-0168-5585, E-mail: yuan352160@126.com。

*本文系“管理科学与工程”省高校人文社科研究基地项目“基于用户节点属性的微博突发话题传播预测算法”(项目编号: GK140203204004/02)和 2015 年杭州电子科技大学研究生优秀学位论文培育基金项目“基于社交媒体的体验性商品销量预测—以票房预测为例”(项目编号: ZX150605304023)的研究成果之一。

出S-PLSA模型，即首先从博客中抽取2 030个评价词，并且使用这些词出现的频率构建特征向量表示每一篇博客，再计算情感倾向，为了提高预测的精度，引入历史销量构建基于情感的自回归模型进行票房预测。但由于新浪微博是短文本，并且含有更多的网络用语以及表情符号，和博客的长文本的情感分析存在很大的差别。Asur 等[10]收集了 24 部影片的推文信息，使用 LingPipe 语言分析包进行情感分类，引入 average-tweet-rate 和正负极性模型构建线性回归模型进行票房预测，实验结果表明，引入正负极性模型的效果更优。然而该方法存在以下缺陷：该文基于"一部电影的正面评论多于负面评论，则该电影的票房可能会更好"这一假设，但在粉丝经济时代，口碑的好坏和电影票房没有直接的关系；线性回归模型中引入的流行度，忽略了不同的情感倾向对票房的贡献度不一样这个事实；由于影响票房的因素很多，使用简单的线性回归无法保证模型的稳定性。Du 等[11]修正了 Asur 等的假设，认为如果有更多的微博内容是和电影票房增长相关，则票房会更好，进而将微博情感分为三类，引入 AS(Authority Score)，通过三种方法进行票房预测，结果发现，神经网络的预测效果最好。但该模型是使用第一周和第二周的历史票房数据来预测第三周的数据，无法实现对电影上映前的预测，同时在情感分类中忽略了领域情感词以及表情符号等对情感分类的影响。

2.2　首映周票房预测

基于微博或博客的票房预测大都引入了历史数据，如使用第一周和第二周的数据进行预测。然而，在电影上映之前却不存在相关的历史数据。因此，如何在没有历史票房数据的情况下进行首映周票房预测是一个巨大的挑战。一种预测方法是利用消费者的观影意愿来预测票房。Eliashberg 等[12]通过问卷调查的方式确定用户的观影意愿，并以此来估计总观影数和票房。Shugan 等[13]通过收集用户在看过预告片之后的观影意愿来预测观影意愿对票房的影响。结果发现，观影意愿和票房之间的相关性较高。

上述关于票房预测相关研究，要么是使用历史数据进行预测，要么是使用小规模问卷调查的方式来搜集用户的观影意愿。但如何利用微博上的信息所汇聚成的集体智慧进行首映周票房预测并为后续周的票房预测提供数据来源仍存在很大的挑战。

通过对微博内容的观察，发现电影主创人员发布的微博大多和预告片、电影海报、主题曲相关，本文利用这些微博评论中的情感所体现出的观影意愿进行首映周的票房预测。因此本文结合情感分析和神经网络提出一种基于微博的首映周票房预测模型。该模型充分利用微博的数量和情感特征，通过支持向量机分类出含有显式意图的评论，将其定义为强正面评论，并在对传统的情感分类准则进行修正后，利用构建的中文微博情感词典对剩余评论进行情感分类，考虑到不同的评论情感类别对票房的不同贡献能力，赋予情感类别以不同的权重，再通过 BP 神经网络进行票房预测。

3　基于微博的电影首映周票房预测

本文研究框架如图 1 所示：

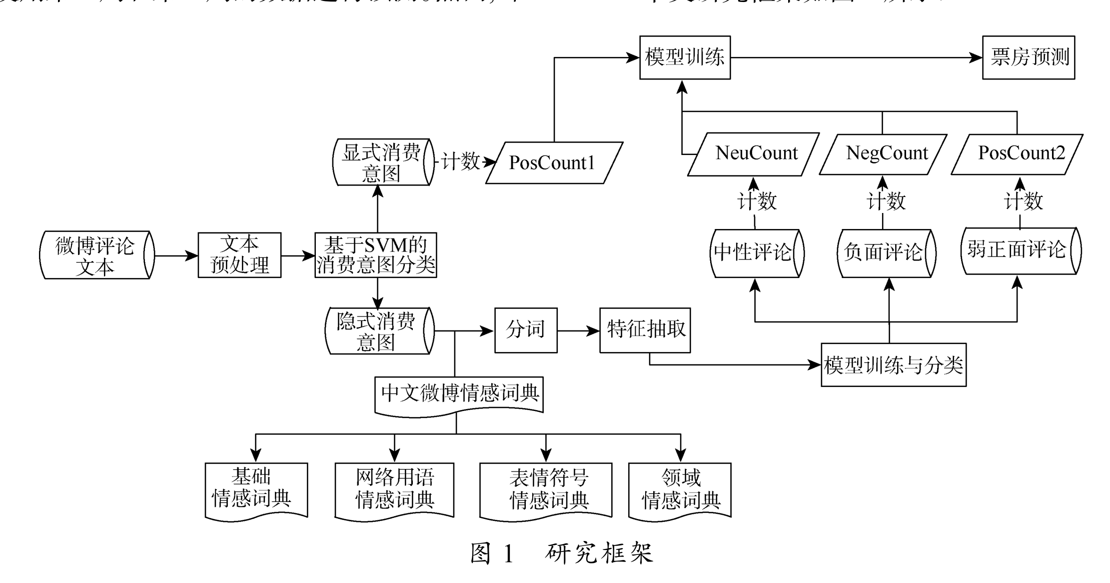


图 1　研究框架

利用微博情感进行首映周票房预测分为6个步骤:

(1) 结合相关文献以及实际的电影市场情况，确定影响首映周票房的因素;

(2) 对微博评论数据进行预处理;

(3) 利用支持向量机从评论中分类出含有显式消费意图的强正面评论;

(4) 构建基于 HowNet 的中文微博情感词典;

(5) 对剩余评论进行分词和情感分类;

(6) 使用步骤(1)中确定的票房影响因素构建预测模型的输入，进行模型的训练和票房预测。

3.1 首映周票房影响因素

微博作为潜在的买家和卖家直接沟通的平台，特别是在电影上映前，电影的主创人员和粉丝互动时，粉丝会发布"是否购买电影票"和"是否有观影意愿"的评论。据统计，用户发布的微博中大约有 3%的微博含有消费意图[14]。而在电影评论领域含有显式消费意图的比例甚至高达 18.7%。与 Eliashberg 等[12]和 Shugan 等[13]的研究一致，本文选择以消费者的观影意愿为基础构建票房预测模型[15]。此外，Asur 等[10]已经证明社交媒体中的情感有助于票房预测。同时 Du 等[11]提出不同的情感倾向对票房的贡献能力是不同的，并且证明该假设比 Asur 等的假设能够更好地预测票房。因此本文提出使用微博中的消费意图、用户评论的情感倾向来进行票房预测。

3.2 文本预处理

由于 XML 文件中存在很多和评论无关的信息，为了提高数据的质量，通过对微博评论进行分析，采用如下方法进行评论的预处理：在 Python 中使用正则表达式 re.compile(r'(?<=<div>)(.*?)(?=</div>)', re.S)获得标签之间的所有微博评论，并存入 MySQL 数据库；去除评论中噪声数据，包括网页链接(http://), "#话题#"和二次转发符号"//"。

3.3 基于 SVM 的消费意图挖掘

(1) 问题描述

消费意图是指购买某个产品的意愿[16]。消费意图识别是指利用各种技术对含有消费需求的数据进行分析，从而识别出用户的消费意图。用户倾向于在社交媒体上隐式或显式地表达他们的购买意愿。例如"井宝，捉妖记票已定好，明天就去看你"。然而，尽管有些用户的帖子提到了和电影相关的信息，但用户并没有在帖子中明确表达他们看电影的意图，例如"大美岩，女神，爱死你了。。。你的电影我一定会看哒"。因此对购买意图的挖掘可以看成是一个二分类问题[17]。为此，对于给定的电影名称，首先要从微博上收集主创人员的微博及其对应的评论，然后对这些评论进行分类，即显式消费意图和隐式消费意图。该部分主要是为了获得已经购票的用户的数量。

(2) 基于 SVM 的消费意图挖掘

在机器学习中, SVM 是由 Vapnik 等[18]引入。SVM 可以用于分类和回归分析。当进行分类时，主要利用核函数将线性不可分的训练数据投射到高维特征空间从而实现线性可分，并且适用于小样本数据的分类。而进行分类时，重要的是模型输入特征的选择，根据对微博评论数据的观察，本文选择以下两个特征。

①提及特征(M)：在微博中，用户使用@来提醒他们的朋友来看这条微博。如"7 月 16 日 煎饼侠，约吗@某人"。通过对微博评论语料的观察，发现大部分包含@的评论包含消费意图。因此，提及特征可以帮助判断一条微博是否包含消费意图。如果某条评论中出现@，则将 M 设为"true"。

②触发词特征(T)：社交媒体用户经常使用一些特殊的词语来表达他们的购买意图，如"已买"来表达其消费意图。本文从评论文本中手工挑选出部分词汇并将这些词命名为触发词。如果一条评论中包含触发词，将 T 设置为"true"。最终本文选择 12 个词作为触发词。

3.4 领域情感词典的构建

对于微博评论文本，除了包含已经购买的用户评论外，还有一部分评论，虽然没有明确表达是否去观影，但用户使用各种表情符号、情感词等表达对电影的看法，该部分评论的用户有可能会转化为最终的观影人员，同时不同的情感倾向对票房的贡献能力是不一样的。因此，为了量化一部电影的情感倾向，本文在构建领域情感词典的基础上进行情感分析。

情感词是最好的表示文本情感的特征之一[19]，丰富的情感词典有助于提升情感倾向性判定效果。文本的情感倾向大多通过情感词语体现，情感词典能否覆盖全面在一定程度上影响着情感分类效果，故情感词典的构建是情感分类研究的基础。本文的中文微博情感词典构成如图 2 所示:

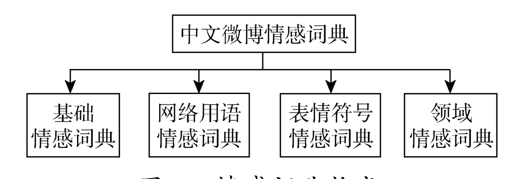


图 2 情感词典构成

(1) 基础情感词典

选用《知网》情感分析用词语集中的正/负面情感词语集(中文)和正/负面评价词语集(中文), 并将其中不常用或是情感倾向在电影评论中有歧义或倾向不明显的剔除掉, 同时还根据电影评论领域的特点对部分情感词的倾向进行修正, 最终得到基础情感词典组成如表 1 所示:

表 1　基础情感词典

词语集名称	数量(个)
正面情感词	4 105
负面情感词	4 234

(2) 网络用语情感词典

网络用语包括两种: 一种是新出现的网络新词, 如"路转粉"; 二是已经存在的词汇, 因被赋予了新的含义而带有一定的情感倾向。而这些网络用语逐渐成为用户表达情感的主力军。因此构建网络用语情感词典显得尤为重要。本文选用"网词网"提供的网络词汇列表并通过赋予不同的词汇相应的情感倾向来构建网络用语情感词典。

(3) 表情符号情感词典

通过对微博的评论观察, 以《煎饼侠》中袁姗姗的微博评论为例, 在 486 条微博评论中有 180 条评论含有微博表情符号, 因此亟需构建一个以微博自带的表情符号为基础的表情符号词典来优化情感分类。其中正面情感标注为"+1", 负面标注为"–1", 结合微博评论语料, 忽略其中有歧义的表情符号, 并根据微博评论对表情符号进行极性的调整, 如"[泪]"一般表示负向情感, 但通过电影评论观察, 发现一般出现在买不到首映票之后, 实际上该条评论传达的是对观影的积极倾向。本文共收集 66 个带有明显情感倾向的表情符号, 其中正面表情符号 43 个, 负面表情符号 23 个。部分如表 2 所示:

表 2　表情符号情感词典示例

情感符号	表达含义	被抓取的表现形式	极性
	泪	[泪]	+1
	甜馨爱你哟	[甜馨爱你哟]	+1
	怒骂	[怒骂]	–1

(4) 领域情感词典的构建

①SO-PMI

点互信息是信息论中度量两个随机变量间统计依赖性的一种测度。在自然语言处理中, 可以用来计算两个词汇 word1 和 word2 之间的相似度。其基本思想是如果 word1 和 word2 在文本中同时出现的概率越大, 则两者的相关性越高, 情感倾向一致。

word1 和 word2 的点互信息 PMI(word1, word2) 计算公式如下:

$$\mathrm{PMI(word1, word2)} = \log_2\left(\frac{\mathrm{P(word1\,\&\,word2)}}{\mathrm{P(word1)P(word2)}}\right) \tag{1}$$

其中, P(word1 & word2)表示 word1 和 word2 同时出现的概率, P(word1)和 P(word2)表示每个词单独出现的概率。

在实际的应用过程中, 公式(1)的概率可以通过 word1 和 word2 在评论语料中出现的次数进行估计, 而出现的次数可以通过使用文档频次法获得。因此公式(1)可以使用如下公式进行近似:

$$\mathrm{PMI(word1, word2)} \approx \log_2\left(\frac{N \times \mathrm{df(word1, word2)}}{\mathrm{df(word1)} \times \mathrm{df(word2)}}\right) \tag{2}$$

其中, N 表示语料库中总的词次数, df(word1) 和 df(word2) 表示两个词在评论语料库中出现的次数, df(word1, word2) 表示两个词在语料库中共同出现的次数。

通过计算 PMI(word1, word2) 的值, 最终根据值所属的范围, 确定 word1 和 word2 之间的关联度, 确定的规则如下:

$$\mathrm{PMI(word1, word2)}\begin{cases} >0 & \text{两个词语相关;值越大,相关性越高} \\ =0 & \text{两个词语是统计独立的,不相关也不互斥} \\ <0 & \text{两个词语不相关,互斥} \end{cases} \tag{3}$$

为了量化两个情感词之间的相似度, 将 PMI 引入到情感分析领域来计算词语的情感倾向(SO), 从而确定候选词的情感倾向。

SO-PMI 的基本思想: 选取基准词集, 包含一组褒义基准词集 Pwords 和一组贬义基准词集 Nwords 。分别计算某个词与 Pwords 和 Nwords 的 PMI , 根据二者的 PMI 差值的取值确定其情感倾向, 以 word 为例, SO-PMI 的计算公式如下:

$$\begin{aligned}\mathrm{SO\text{-}PMI(word)} &= \sum_{\mathrm{Pword}\in\mathrm{Pwords}} \mathrm{PMI(word, Pword)} - \sum_{\mathrm{Nword}\in\mathrm{Nwords}} \mathrm{PMI(word, Nword)} \\ &\approx \log_2 \frac{\sum_{\mathrm{Pword}\in\mathrm{Pwords}} \left(\frac{N\times \mathrm{df(word, Pword)}}{\mathrm{df(word)}\times \mathrm{df(Pword)}}\right)}{\sum_{\mathrm{Nword}\in\mathrm{Nwords}} \left(\frac{N\times \mathrm{df(word, Nword)}}{\mathrm{df(word)}\times \mathrm{df(Nword)}}\right)}\end{aligned} \tag{4}$$

通过比较 $\sum_{\mathrm{Pword}\in\mathrm{Pwords}} \mathrm{PMI(word, Pword)}$ 和 $\sum_{\mathrm{Nword}\in\mathrm{Nwords}} \mathrm{PMI(word, Nword)}$ 的大小, 可以确定 word 的情感倾向并将其加入相应的情感词典。一般将 0 作为阈值, 根据与 0 的大小关系, 共有三种不同的情况:

$$\mathrm{SO\text{-}PMI(word)}\begin{cases} >0 & \text{褒义词} \\ =0 & \text{中性词} \\ <0 & \text{贬义词} \end{cases} \tag{5}$$

②领域情感词典的生成

领域情感词典的生成包括两个步骤：基准词的选取；使用 SO-PMI 判断候选词是否是情感词及其情感倾向。

1) 基准词的选取

郭叶[20]通过对比 10 对、20 对、30 对、40 对基准词对情感倾向判断的准确率，结果发现选取 40 对基准词时，准确率高达 81.37%。本文对 115 202 条微博评论语料进行词频统计，按照由高到低进行排序，人工挑选出出现频率高并且情感倾向明显的词语作为基准词，其中褒义基准词 20 个，贬义基准词 20 个。

2) 基于 SO-PMI 的候选词识别

对 115 202 条微博评论，使用候选词识别算法 SO-PMI 构建领域情感词典，最终获得褒义词 732 个，贬义词 507 个。具体的算法流程如图 3 所示：

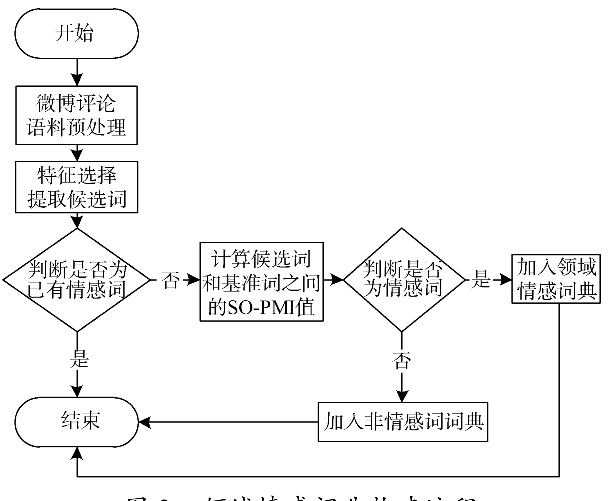


图 3　领域情感词典构建流程

3.5　情感分类

(1) 研究假设及规则定义

本文使用 Du 等[11]提出的假设，但对不同的情感倾向内容进行修正：

假设：如果有更多的微博评论是和票房增长相关，则电影的票房会更好。基于这个假设，笔者将经过主客观分类的微博评论分为三类，具体的分类准则如下：

规则1：正面评论，包括强正面评论和弱正面评论。

规则1.1：强正面评论：即表达“已经购票”的用户评论。

例如：已经提早买过票啦，就等上映 le[鼓掌]；凌晨就不能去了，但是已经买了明天晚上的票，一定要看美美的你！

规则 1.2：弱正面评论：对影片、演员等表达积极的情感。

例如：大美岩[给劲]票房大麦；其实蛮欣赏大鹏的电影上映会去支持。

规则 2：中性评论：是指对影片、演员、情节等表达中性情感。

例如：井柏然是相爱穿梭前年的男主角啊。

规则 3：负面评论：是对影片、演员等表达消极、负面的情感。

例如：炒作过头了，本来想看都不想去了。

基于上述假设，本文提出用户影响力特征(MU)，定义如下：

$$MI(i)=\alpha_1\times PosCount1+\alpha_2\times PosCount2+\beta\times NegCount+\gamma\times NeuCount \tag{6}$$

MI(i) 是第 i 条微博的影响力，PosCount1 是强正面评论数，PosCount2 是弱正面评论数，NegCount 是负面评论数，NeuCount 是中性评论数，α_1、α_2、β、γ 是相应的权重，可通过样本数据训练获得。

因此一个用户的影响力为：

$$MU(j)=\frac{\sum_{i=1}^{N}MI(i)}{N} \tag{7}$$

其中，N 为该用户所发布的微博数量。

为了使用用户评论中的情感能够同时更好地用于票房预测，本文定义 X_m 作为样本的输入特征：

$$X_m=\sum_{i=1}^{m}MU(j) \tag{8}$$

其中，m 是和该电影相关的微博用户数，本文定义 m = 4，只选择 4 位主创人员的原因在于王铮等[21]使用前三位主演的名气作为票房的其中一个变量，其对票房的解释能力较高，并且发现随着演员数量的增加，片酬的增加，会存在挤出效应，从而使票房降低，因此本文选择基于王铮等[21]的研究，并结合百度百科对一部电影的演员的介绍规则，选择前 4 位主演的微博评论，包括男一、女一、男二和女二。

(2) 情感分类

由于 140 字的微博信息量少，可区分度低，因此，在以词为维度的向量空间模型中，呈现出特征稀疏的特点。因此需要构建电影领域的情感词典来降低稀疏性，并利用构建的情感词典进行情感分类，具体算法流程如下：

输入：微博剩余评论 $D=(d_1,d_2,\cdots,d_k,\cdots)$，情感词典 Dic。

输出：评论分类结果，具体包括弱正面评论个数PosCount2，中性评论个数NeuCount，负面评论个数NegCount。

①使用NLPIR[①]汉语分词系统对$\forall d_k (d_k \in D)$进行分词并去掉停用词。

②在游建平[22]提出的上下文滑动算法基础上，将词性规则、情感词典、平滑算法相结合，对微博评论的情感相关特征项进行抽取。本文需要抽取的特征包括情感词、程度副词、否定词、表情符号、网络用语。

③使用LIBSVM进行分类器的训练，并完成对评论文本的情感分类。

3.6 基于BP神经网络的票房预测

BP算法是一种有监督的学习算法，是使用反向传播算法对网络的权值和偏差进行反复调整训练的一种多层前馈神经网络。BP神经网络分为两个过程：信号的正向传播；误差的反向传播。在BP神经网络中，单个样本有m个输入，有n个输出，在输入层和输出层之间通常还有若干个隐含层。由于输入层和输出层的节点个数都是确定的，而隐含层节点个数不确定，根据经验公式$h=\sqrt{m+n}+a$ (a为1-10之间的调节常数)确定隐含层的个数。本文使用R语言中的nnet包进行票房预测。nnet包提供了反向传播算法并且可以自定义设置激活函数。实现步骤如下：

输入：电影的正面评论(强正面评论和弱正面评论)、负面评论和中性评论的个数。

输出：票房的预测值和平均绝对百分比误差(MAPE)。

①读入神经网络nnet包和MySQL数据库连接时需要使用的RODBC包；

②通过定义odbcConnect与数据库建立连接，利用sqlFetch获得微博评论情感分类结果；

③按照公式(6)至公式(8)计算获得X_m；

④使用splitForTrainingAndTest划分测试集和训练集；

⑤使用scale对数据进行标准化；

⑥使用nnet命令，参数规定隐含层的个数、学习率η_1，最大迭代次数；

⑦使用predict进行预测；

⑧计算MAPE。

4 实验设计与结果

4.1 数据集

本文使用的数据集由两部分组成：票房数据和电影微博评论文本。票房数据来自中国电影票房数据中心，包括《捉妖记》、《同桌的你》、《煎饼侠》等10部电影的首映周票房。微博评论文本是通过网络爬虫软件GooSeeker在新浪微博上进行采集。获取的评论文本是电影主创人员在电影上映前一周发布的和电影相关的微博及其评论。最终，共收集到115 202条微博评论数据。

为了证明本文方法的有效性，首先指导用户对语料进行标注。两个注释人员对数据集中的每一条微博评论标注是否包含购买意图。对前两个注释人员存在争论的微博，由第三个注释人员评判其所属的类别。消费意图标注集共有3 000条评论文本，其中2 300条是训练集，700条是测试集。同时为了比较情感分类的准确率，每部电影各选择2 000条评论进行人工标注，选择5名志愿者分别对20 000条微博评论进行情感倾向判断。最后，对5个志愿者的标注结果进行汇总，以每条评论中情感倾向得票数最多的为准。消费意图和情感分类人工标注结果如表3所示：

表3 人工标注结果

电影	强正面评论	弱正面评论	中性评论	负面评论
捉妖记	296	1 671	0	33
小时代3	319	1 305	14	362
何以笙箫默	279	1 592	18	111
小时代4	307	1 295	0	398
同桌的你	274	1 408	94	224
后会无期	238	1 386	347	29
匆匆那年	315	1 509	9	167
煎饼侠	519	1 463	0	18
夏洛特烦恼	406	1 496	81	17
栀子花开	361	1 107	151	381

4.2 评价指标

(1) 分类模型评价指标

使用准确率(Precision)、召回率(Recall)和F1作为分类效果评估指标。

$$\text{Precision}=\frac{\text{被分类器判定的正例中真正的正例}}{\text{被分类器判定的正例}} \tag{9}$$

$$\text{Recall}=\frac{\text{被正确判定的正例}}{\text{总的正例}} \tag{10}$$

①http://ictclas.nlpir.org/.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

(2) 票房预测模型评价指标

平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)是通过计算神经网络输出的预测值与其实际票房数据之间的误差后，计算误差与实际值的比值的绝对值。MAPE 越小，代表预测效果越好。

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \quad (12)$$

4.3　Baseline 实验

Asur 等[10]不仅证实了社交媒体具有预测现实世界的能力，还使用线性回归模型并从 Tweet 中抽取 tweet-rate 和 Pnratio 特征成功地预测了首映周票房。在本文的数据集上利用 Asur 等的方法进行票房预测，并作为 Baseline 实验。

4.4　实验结果与分析

(1) 消费意图挖掘实验结果分析

为了验证本文进行消费者意图挖掘时选取的两类特征的有效性，采用逐个增加特征的方法来验证。首先只使用提及特征(M)作为分类特征，然后增加触发词特征(T)，最终的实验结果如表 4 所示:

表 4　特征选择的实验结果

特征	Precision	Recall	F1
提及特征(M)	0.45	0.67	0.56
提及特征(M)+触发词特征(T)	0.73	0.81	0.77

和人工标注的强正面评论相比，表 4 的数据表明通过向分类器中增加特征可以改善消费意图的分类结果。同时也证实了选择的两个特征可以有效地从评论文本中区分出显式消费意图。此外从数据中可以看出，触发词特征对分类器效果的改进很大。

(2) 情感分类实验结果分析

以电影《捉妖记》、《煎饼侠》和《小时代 4》为例说明使用本文构建的领域情感词典情感分类的准确率，从表 5 可以看出，以人工标注作为基准，使用本文构建的领域情感词典具有较高的情感分类准确率。

表 5　情感分类实验结果

电影 / 评价指标	捉妖记	煎饼侠	小时代 4
Precison	0.8973	0.9104	0.8274

(3) 情感分类标准对票房的影响

传统的情感分类仅仅将对电影情节等具有正面评价的评论归入正面评论，但忽略了在现有的预售制环境下用户会提前订票这一事实，而订票的这一类用户和仅仅对电影表达情感倾向的用户相比对票房的贡献能力更大。因此使用线性回归模型对首映周票房进行回归预测，将本文的情感分类准则获取的特征作为输入并和 Asur 等[10]的情感分类标准相比，如表 6 所示:

表 6　不同的特征和票房的相关系数

特征	相关系数
Asur 等的情感分类准则	0.2906481
本文的情感分类准则	0.3194738

从表 6 可以看出，使用本文修正后的情感分类准则获得的特征和票房的相关度更高。同时为了验证使用神经网络的预测效果，将上述特征作为神经网络的样本输入，随机选择两部电影作为训练集，剩余的 8 部电影作为测试集，使用 MAPE 作为评价指标。具体结果如表 7 所示:

表 7　不同特征的样本输入的平均 MAPE 值

情感分类	MAPE
本文的情感分类准则	0.1978201
传统的情感分类准则	0.2013058

通过表 7 发现，使用本文提出的分类准则计算的 X_m 作为样本特征输入进行票房预测的结果降低了 1.73 个平均绝对百分比误差。为了验证 X_m 中不同的特征对模型的解释能力，依次使用 Tweet、Pnratio、强正面评论和弱正面评论进行票房预测。图 4 说明消费意图可以更好用于改善票房预测，并且优于评论文本中的情感倾向对票房的预测能力。

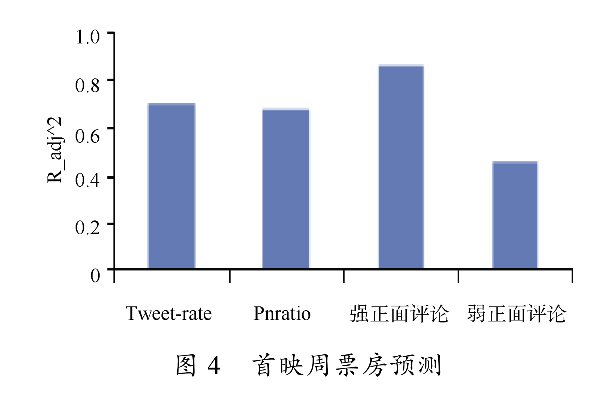


图 4　首映周票房预测

(4) 不同预测模型比较

现有的票房预测模型一般使用神经网络、SVM 或 LR，而本文使用神经网络进行预测，并对比了不同模型的预测效果。表 8 中的数据表明，使用 BP 神经网络的效果要明显优于其他两种方法。

表 8　不同预测效果的模型比较

预测模型	MAPE
BP 神经网络	0.1912093
SVM	0.2173051
LR	0.2640637

相比于线性回归模型，非线性回归模型有更好的预测效果。可能是由于影响票房的因素较多，而线性回归模型的稳定性较差。

5　结　语

本文从微博评论中识别出显式消费意图，在对传统的文本情感分类进行修正的基础上，定义了基于微博情感分析的用户影响力 MU，并且对不同的情感倾向赋予不同的权重。为了更好地实现情感分类效果，构建了基于电影领域的情感词典，并且证实该词典能够更好地用于情感分类。最后对比了不同的预测模型的票房预测效果。

但本文仍存在以下不足：仅仅对首映周的票房进行预测，没有建立一个能够实现对后续周票房进行预测的动态模型；由于语料不充分，构建的中文微博情感词典，可能无法在所有的电影微博评论中表现出较好的分类效果。

参考文献：

[1] 尹裴，王洪伟，郭恺强．中文产品评论的“特征观点对”识别：基于领域本体的建模方法[J].系统工程，2013，31(1): 68-77. (Yin Pei, Wang Hongwei, Guo Kaiqiang. Feature-opinion Pair Identification in Chinese Online Reviews Based on Domain Ontology Modeling Method [J]. Systems Engineering, 2013, 31(1): 68-77.)

[2] 张闯，姜杨，吴铭，等．基于社会化媒体节点属性的信息预测[J]．北京邮电大学学报，2012，35(4)：24-27. (Zhang Chuang, Jiang Yang, Wu Ming, et al. Information Predictions Based on Node Attributes of Social Media [J]. Journal of Beijing University of Posts and Telecommunications, 2012, 35(4): 24-27.)

[3] Liu Y, Huang X, An A, et al. ARSA: A Sentiment-aware Model for Predicting Sales Performance Using Blogs [C]. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007: 607-614.

[4] Neelamegham R, Chintagunta P. A Bayesian Model to Forecast New Product Performance in Domestic and International Markets [J]. Marketing Science, 1999, 18(2): 115-136.

[5] Elberse A, Eliashberg J. Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures [J]. Marketing Science, 2003, 22(3): 329-354.

[6] Liu Y. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue [J]. Journal of Marketing, 2006, 70(3): 74-89.

[7] Sawhney M S, Eliashberg J. A Parsimonious Model for Forecasting Gross Box-office Revenues of Motion Pictures [J]. Marketing Science, 1996, 15(2): 113-131.

[8] Eliashberg J, Shugan S M. Film Critics: Influencers or Predictors? [J]. Journal of Marketing, 1997, 61(2): 68-78.

[9] Krider R E, Weinberg C B. Competitive Dynamics and the Introduction of New Products: The Motion Picture Timing Game[J]. Journal of Marketing Research, 1998, 35(1): 1-15.

[10] Asur S, Huberman B A. Predicting the Future with Social Media [C]. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2010: 492-499.

[11] Du J, Xu H, Huang X. Box Office Prediction Based on Microblog [J]. Expert Systems with Applications, 2014, 41(4): 1680-1689.

[12] Eliashberg J, Jonker J J, Sawhney M S, et al. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures[J]. Marketing Science, 2015, 19(3): 226-243.

[13] Shugan S M, Swait J. Enabling Movie Design and Cumulative Box Office Predictions Using Historical Data and Consumer Intent-to-View [R]. University of Florida, 2000.

[14] 陈晓东．基于情感词典的中文微博情感倾向分析研究[D]．武汉：华中科技大学，2012.(Chen Xiaodong. Research on Sentiment Dictionary Based Emotional Tendency Analysis of Chinese Microblog [D]. Wuhan: Huazhong University of Science & Tchnology, 2012.)

[15] 王炼，贾建民．基于网络搜索的票房预测模型——来自中国电影市场的证据[J]．系统工程理论与实践，2014，34(12): 3079-3090. (Wang Lian, Jia Jianmin. Forecasting Box Office

Performance Based on Online Search:Evidence from Chinese Movie Industry [J]. Systems Engineering-Theory & Practice, 2014, 34(12): 3079-3090.)

[16] Fu B, Liu T. Weakly-supervised Consumption Intent Detection in Microblogs [J]. Journal of Computational Information Systems, 2013, 6(9): 2423-2431.

[17] 陈浩辰. 基于微博的消费意图挖掘[D]. 哈尔滨: 哈尔滨工业大学, 2014.(Chen Haochen.Consumption Intention Mining Based on Microblog [D]. Harbin: Harbin Institute of Technology, 2014.)

[18] Vapnik V N, Vapnik V. Statistical Learning Theory [M]. New York: Wiley, 1998.

[19] 杜伟夫, 谭松波, 云晓春, 等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展, 2009, 46(10): 1713-1720. (Du Weifu, Tan Songbo, Yun Xiaochun, et al. A New Method to Compute Semantic Orientation[J]. Journal of Computer Research and Development, 2009, 46(10): 1713-1720.)

[20] 郭叶. 中文句子情感倾向分析[D]. 北京: 北京邮电大学, 2010.(Guo Ye. Sentiment Orientation Analysis of Chinese Sentences [D]. Beijing: Beijing University of Posts and Telecommunications, 2010.)

[21] 王铮, 许敏. 电影票房的影响因素分析——基于 Logit 模型的研究[J]. 经济问题探索, 2013 (11): 96-102. (Wang Zheng, Xu Min. Analysis of the Influence Factors of Movie Box Office—Based on Logit Model [J]. Inquiry into Economic Issues, 2013 (11): 96-102.)

[22] 游建平. 基于语义情感空间模型的微博情感倾向性研究[D]. 广州: 暨南大学, 2012. (You Jianping. Micro-Blog Sentiment Analysis Based on Semantic Sentiment Space Model [D]. Guangzhou: Jinan University, 2012.)

作者贡献声明:

王晓耘, 袁媛: 提出研究思路, 设计研究方案;
袁媛, 史玲玲: 采集和分析数据, 进行实验;
袁媛: 论文起草;
王晓耘: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 袁媛. 原始数据.rar. 电影主创人员的微博评论(10 部电影).

收稿日期: 2015-09-11
收修改稿日期: 2016-01-08

Predicting Opening Weekend Box Office Prediction Based on Microblog

Wang Xiaoyun Yuan Yuan Shi Lingling
(Management School, Hangzhou Dianzi University, Hangzhou 310012, China)

Abstract: [**Objective**] This study aims to solve the problems of the existing pre-release box office prediction models due to data constraints and other factors. [**Methods**] We first retrieved microblog comments, and then used SVM to identify explicit consumer intention, namely strong positive comments. Second, we modified the traditional sentiment classification schemes to build a Chinese microblog sentiment dictionary based on HowNet. Finally, we defined a new user influence feature and used the BP neural network to predict box office. [**Results**] The proposed model could forecast the opening box office more accuately. [**Limitations**] Due to inadequate corpus, the sentiment dictionary may not work well for all microblog movie comments. A dynamic forecasting model was not established between the pre-release and post-release period. [**Conclusions**] The proposed model can effectively predict opening box office.

Keywords: Sentiment dictionary Sentiment classification Opening weekend box office prediction Neural network